12-2008

On visualizing heterogeneous semantic networks from multiple data sources

Maureen Maureen
Nanyang Technological University

Aixin SUN
Nanyang Technological University

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Anwitaman DATTA
Nanyang Technological University

Kuiyu CHANG
Nanyang Technological University

DOI: https://doi.org/10.1007/978-3-540-89533-6_27

Published in Digital Libraries: Universal and Ubiquitous Access to Information: 11th International Conference on Asian Digital Libraries, ICADL 2008, Bali, Indonesia, December 2-5, 2008. Proceedings. pp. 266-275.
http://dx.doi.org/10.1007/978-3-540-89533-6_27

On Visualizing Heterogeneous Semantic Networks from Multiple Data Sources

Maureen[1], Aixin Sun[1], Ee-Peng Lim[2], Anwitaman Datta[1], and Kuiyu Chang[1]

[1] School of Computer Engineering, Nanyang Technological University, Singapore
{maureen, axsun, anwitaman, askychang}@ntu.edu.sg
[2] School of Information Systems, Singapore Management University, Singapore
eplim@smu.edu.sg

Abstract. In this paper, we focus on the visualization of heterogeneous semantic networks obtained from multiple data sources. A semantic network comprising a set of entities and relationships is often used for representing knowledge derived from textual data or database records. Although the semantic networks created for the same domain at different data sources may cover a similar set of entities, these networks could also be very different because of naming conventions, coverage, view points, and other reasons. Since digital libraries often contain data from multiple sources, we propose a visualization tool to integrate and analyze the differences among multiple social networks. Through a case study on two terrorism-related semantic networks derived from Wikipedia and Terrorism Knowledge Base (TKB) respectively, the effectiveness of our proposed visualization tool is demonstrated.

1 Introduction

1.1 Motivation

A semantic network refers to a set of concepts or entities, possibly of different types, connected by relationships. In the digital library context, semantic networks have always been a useful representation for representing knowledge found in text and database records which in turn helps users to more effectively and quickly search and navigate information. Some often cited examples of semantic networks in digital libraries include author co-citation networks [2], keyword co-occurrence networks [10], etc. In this paper, we focus on social networks as kinds of semantic networks found in text collections and databases. For large social networks, visualization tools will be required to assist users in viewing, searching and analyzing entities and relationships in the networks as well as locating the documents or database records containing the sub-networks users are interested in. In this paper, we therefore describe our proposed interactive tool that supports social network visualization and data access based on network navigation.

As digital libraries often include data taken from different data sources, the social networks obtained from one source may look very different from other

sources even when they share some common entities and relationships. This heterogeneity is often caused by different *naming conventions*, *attribute format*, *coverage*, and *view points* adopted at different sources. For example, the *(first name, last name)* person name format may be used in source A, while source B uses the *(last name, first name)* name format. Person entities from A may have a phone attribute but not those from B. As the social networks can be contributed by different sets of users, they may not cover the same set of entities and relationships. Furthermore, the users responsible for creating content at different sources may assign different type labels or attribute values to the same entity or relationship due to their distinctive view points. Given these heterogeneity issues, a visualization tool for such social networks will be required to integrate multiple social networks together via entity (and relationship) resolution as well as attribute merging and to keep the unresolved and resolved entities distinctive in the user interface.

With the recent advances in social computing and the wide availability of social software (e.g., wikis and blogs), it is increasingly easy to find semantic networks or even social networks of specific domains defined over Web content or publicly accessible databases. For example, Wikipedia, the largest encyclopedia on the Web, collaboratively created by millions of users provides rich article content about entities which are linked to one another thereby providing additional semantics about their relationships (e.g., topic category labels of articles).

1.2 Objective and Contribution

The main objective of this research is to develop an interactive tool for visualizing semantic networks from multiple data sources. Other than viewing and navigating network entities and relationships, the visualization tool will assist users in exploring the underlying data (documents or database records) from which the networks are obtained, and comparing the entities, relationships, and network connectivities between semantic networks.

Figure 1 depicts the system architecture of the visualization tool. It consists of a **network extractor** that extracts semantic networks from data sources. The extracted network information is stored in the **network database**. The **network integrator** is responsible for taking two or more heterogeneous semantic networks and integrating their entities and relationships. These integrated semantic networks are then be stored in the network database. The **network viewer** provides an interactive interface for users to retrieval semantic networks, navigate them and access semantic networks and their underlying text or database records.

In this paper, we describe our visualization tool built based on the above system architecture and summarize the research contributions as follows:

- We have defined a database schema for modeling semantic networks and the entity matchings between semantic networks. This database schema is designed to be generic enough to handle as many different types of semantic networks as possible.

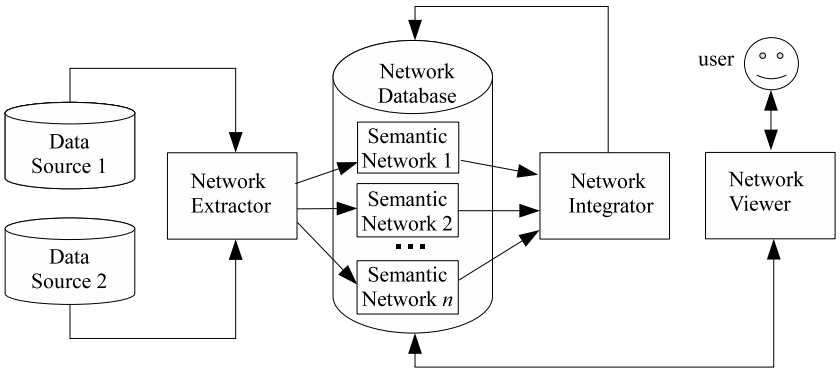

Fig. 1. System Architecture for Visualization Tool

- We have developed a working prototype visualization tool using TouchGraph API [14], a graphical user interface programming package for graph visualization. We use color and shape to distinguish the different data sources and entity types.
- We have applied our tool to a case study involving two terrorism related social networks from (a) Wikipedia and (b) Terrorism Knowledge Base (TKB). TKB is provided on the Web and maintained by the Memorial Institute for the Prevention of Terrorism (MIPT). In this case study, the social network derived from Wikipedia represents the common web user knowledge in the terrorism domain as users acquire information from news articles and other online sources (some of them are mentioned as references in Wikipedia articles). TKB on the other hand is an expert maintained knowledge base containing information about terrorist groups and members. This case study leads to some interesting observations of the integrated social networks, which help users identifying discrepancies between TKB and Wikipedia social networks.

1.3 Paper Outline

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 discusses the modeling and the integrating of semantic networks. The visualization interface is given in Section 4 followed by a case study in Section 5. Finally, Section 6 concludes this paper.

2 Related Work

There has been several works on the visualization of different kinds of network graphs. For example, Vizster provides visualization functions for exploration,

search and analysis of online social networks [5]. Gene network visualization is addressed in [3]. A survey of visualization techniques for ontology networks is reported in [8]. All of the aforementioned network visualization tools can not handle multiple networks as they are confined to display only single networks. In our research, we propose the idea of visualizing multiple semantic networks by integrating them together. The integrated network allows us to better understand the global network connectivities and compare the network differences.

With reference to the survey by Katifori *et al* in [8], our visualization tool adopts both *context plus focus* and *distortion* techniques. Our semantic network graphs when displayed in the network viewer require a combination of context and focus. That is, each graph has a node serving the focus (central node) surrounded by other nodes with edges connected to it. Some of the other visualization tools that use this technique include TGVizTab [1], MoireGraphs [7], OntoRama [4], OntoViz [11],and OZONE [13] .

A closely related work to our visualization tool is Protege [12], a framework for ontology creation, editing and visualization. Under this Protege framework, several visualization tools have been developed including the above-mentioned tool such as TGVizTab and OntoViz. Our tool is similar in the graph visualization aspect but differs in usage and data storage aspects. In particular, we aim to use our visualization tool for multi-modal social networks which are stored in a relational database.

3 Modeling and Integration of Semantic Networks

3.1 Semantic Network Representation

A semantic network consists of typed entities and relationships. Our network data model supports a configurable set of *entity types* and *relationship types*. Each entity type defines a set of attributes shared by all entities belonging to the type; each entity type may have one or more *relationship type* with other entity types. For example, in our case study, the semantic network created in the terrorism domain involves two entity types: Terrorist Group and Terrorist. Terrorist Group entity type has attributes: name, location, and date. Terrorist Group may be related to itself by a Associated With relationship type, and to Terrorist entity type by an Has Leader relationship type. At the instance level, the Terrorist Group entity *Al-Qaeda* has a Associated With relationship with *Yemen Islamic Jihad* (an entity of Terrorist Group) and a Has Leader relationship with *Osama bin Laden*, a Terrorist entity. In our visualization, like many others, each entity is depicted as a *node*, and each relationship is depicted as an directed *edge* connecting the related pair of nodes.

To store the semantic networks from different data sources in our network database, we define meta-data to describe the data sources and their mappings. Each data source is an instance of the Source entity type, identified by its *SourceID*. Each data source is also given a *SourceName* and it consists of one or more *EntityType* instances. *EntityType* instances can be related to other *EntityType*

instances through some relationships. Other than *EntityTypeID* and *EntityTypeName*, each *EntityType* instance may have attributes defined through *Attribute* instances. Specifically, each *Attribute* instance is given a *AttributeName*, *Order* and *IsMultivalued* flag. The *Order* value indicates the relative position at which the attribute will be subsequently displayed by the network viewer. The *IsMultivalued* flag is a boolean value indicating whether the attribute allows set values. An *Attribute* instance is also assigned a *Domain* instance which defines the *DataType*, *MinValue* and *MaxValue* of the the attribute value. Our default *Domain* instances include integer, character strings, date, and float numbers, which are supported by most existing database systems. A *Domain* instance which is user defined will have its enumerable domain values given by the multivalued attribute *UserDefined*. As in many database systems, by separating domain information from the attribute definition, the same *Domain* instance can be shared among different *Attribute* instances. The *EquivalentTo* relationship is used to store matching entities discussed in the following subsections.

3.2 Semantic Network Integration

To integrate different heterogeneous semantic networks, the mapping of entities and relationships between networks need to be addressed. There are two kinds of entity matching, namely *inter-source* and *intra-source* entity matchings. The former refers to finding matching entities from different data sources, while the latter detects matching entities from the same data source.

Inter-Source Entity Matching In this kind of entity matching, we aim to find common real-world entity with different names from different data sources, i.e., *synonyms*. When the difference between two synonyms is minor, they can be detected by a simple name similarity test. An example of this is a terrorist group known as *Harakat-ul-jihad-i-islami* and *Harakat-ul-jihad-ul-islami* defined in TKB and Wikipedia respectively. We measure the similarity between them using edit distance which is the minimum number of operations (character insertion, deletion, or substitution) required to transform one name to another. When the edit distance between two entity names is smaller than a specified threshold (30% of the shortest name length in our case study), we flag entities as candidate synonyms for subsequent human verification. Fuzzy search provided by Lucene is utilized in our implementation to automate the above matching process. However, for synonyms that are very different, name similarity test fail due to their low similarity score. For example, a terrorist group known as *Black Widow* in TKB is known as *Shahdika* in Wikipedia. One can only tell they are synonyms by reading the content of the Wikipedia article and the corresponding TKB database record, as well as referring to external knowledge. For such kind of synonyms, manual matching is adopted in our current implementation. To reduce the amount of manual effort, we will only match entities that have not been matched earlier to some other entities via the name similarity test.

Table 1. Entities in TKB matching ASALA in Wikipedia

1. Armenian Secret Army for the Liberation of Armenia (ASALA)
2. Third of October Group
3. Ninth of June Organization
4. New Armenian Resistance (NAR)
5. September-France

Intra-Source Entity Matching Each real world entity is supposed to to be represented by a unique entry in a data source. However, this assumption does not always hold as the same entity may be labeled differently in a single data source. Some data sources may store these different names of the same entity and their mappings within their databases or markup articles. We propose an intra-source entity matching scheme that derive matching of entity names from the same source by referring to matching entity names in other sources. For example, in Table 1, all the five groups in TKB match a single Wikipedia article called *Armenian Secret Army for the Liberation of Armenia* (ASALA). The reason is that TKB lists the groups: (*Third of October Group*, *Ninth of June Organization*, *New Armenian Resistance*, and *September France*) as possible sub-groups or ad-hoc groups of the more established group named ASALA. These mappings of different names to the same entity can be applied to find matching entity names in Wikipedia.

4 Network Visualization Interface

As shown in the system architecture, our network viewer provides visualization functions for semantic networks stored in a network database. The main visualization functions include: (a) loading and displaying multiple semantic networks; (b) browsing the attributes of nodes; and (c) constructing a subnetwork as part of data analysis. The visualization interface has been implemented using TouchGraph [14], an open-source library in Java for creating and displaying networks through interactive user interface.

4.1 Interface Design

The user interface of network viewer is shown in Figure 2. A drop-down-list at the top section provides users a list of entities to be selected for analysis. Once an entity is selected, its entity profile and attribute information will be displayed at the right section. Since an entity may appear in multiple data sources, the selected entity's information is obtained from all data sources containing it and is shown in the respective source's tabbed pane. The *balloon graph view* [6] is chosen in Touchgraph to display a semantic network containing the selected entity at the center of the network.

We use *color* and *shape* to distinguish the data source(s) and the entity type respectively. In the example given, exclusive information from TKB are in

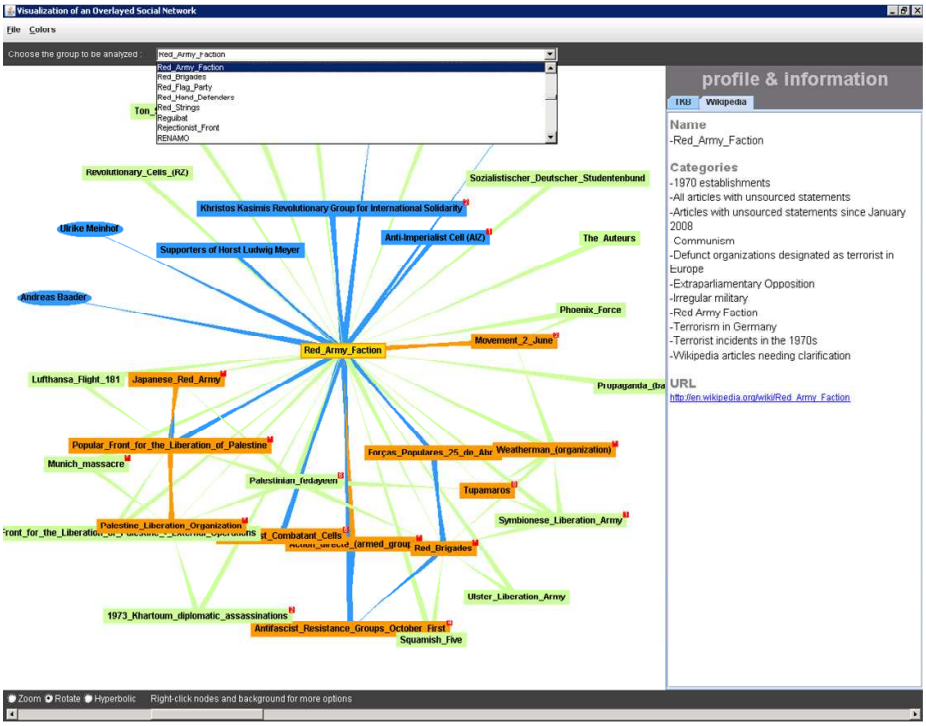

Fig. 2. Look and feel of our visualization tool

blue. Green is assigned to exclusive Wikipedia, and orange is assigned to the overlapping sources (i.e., those that appear in both TKB and Wikipedia). Note that, the color scheme can be configured by users. Moreover, all Terrorist entities are shown in ellipses and Terrorist Group entities are in rectangles. For instance, as shown in Figure 2, an entity named *Andreas Baader* belongs to the Terrorist entity type from TKB. The corresponding node is an ellipse and is in blue. Another entity named *Red Army Faction* belongs to Terrorist Group and can be found in both TKB and Wikipedia. The corresponding node is a rectangle and is in orange. Tools including zooming, rotating, etc are provided at the bottom of the interface.

4.2 Database Configuration

Other than visualizing semantic networks, our visualization tool also supports configuration of the data sources, entity types and their attributes to minimize the user effort in maintaining the databases. This user interface is a wizard-dialog that can (a) add new data source to the network database, and (b) create new entity types. Screen captures are not shown due to the page limit. All the above operations affect the network database content. As soon as a user

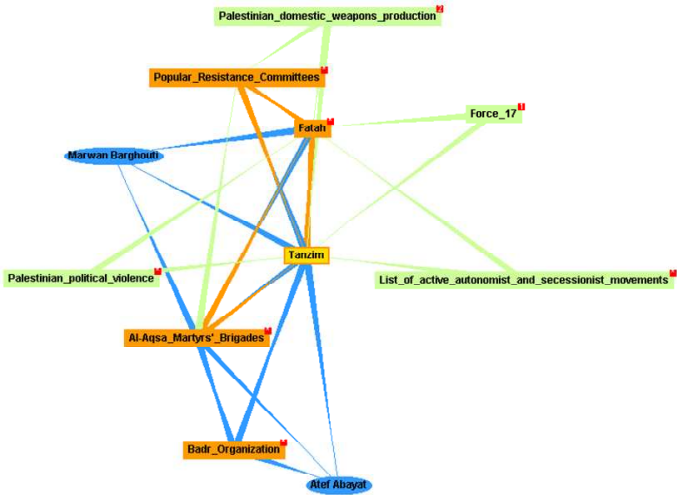

Fig. 3. Network of Tanzim based on information from TKB and Wikipedia

completes configuration using this wizard, the necessary tables in the network database will be automatically built and/or updated. Users may then import, view, insert, edit, remove, and export network data in the network database. To allow semantic network data to be portable across applications, we adopted eXtensible Markup Language (XML) for data import and export operations [9]. These functions are provided mainly for those users who are less familiar with database systems.

5 Case Study

In this section, we demonstrate the usefulness of the graph for social network analysis through a case study. Following our earlier discussion, our case study involves two semantic networks both consisting of terrorism related entities and relationships from TKB and Wikipedia respectively. The semantic network derived from Wikipedia represents the common web user knowledge in the terrorism domain while the one from TKB represents the expert understanding of the domain. Here, we would like to find out how the knowledge of experts differ from that of the public.

For Terrorist Group, 858 entities and 1179 relationships were extracted from TKB; 998 entities and 2302 relationships from Wikipedia. Among them 305 entities and 259 relationships appear in both source. For Terrorist entity type, 1463 entities have been extracted from TKB together with 1374 relationships between Terrorist and Terrorist Group. For Wikipedia, since there is no particular category label for extracting terrorists, extracting terrorists from Wikipedia remains challenging. In this case study, we hence mainly focus on the differences among terrorist groups. As shown in Figure 3, the selected terrorist group *Tanzim* is

shown at the center of the network. Those nodes that only appear in one data source are clearly indicated by their colors. Recall that all information derived from TKB are shown in blue and that from Wikipedia in green; and orange is used for information that derived from both sources. It is therefore interesting to observation differences in relationships among entities that appear in both data sources. For example, according to TKB, *Tanzim* is related to *Badr Organization*, *Al-Aqsa Martyr's Brigades*, *Fatah*, and *Popular Resistance Committee*. On the other hand, according to Wikipedia *Tanzim* is related to all these groups except *Badr Organization*. Furthermore, there are no relationship from *Badr Organization* to *Al-Aqsa Martyr's Brigades* in Wikipedia whereas in TKB such relationship exists. Also, we have observed that according to Wikipedia, there is a relationship between *Popular Resistance Committee* and *Al-Aqsa Martyr's Brigades* whereas it is not mentioned in TKB. This specific example illustrates that in the homeland security domain, the knowledge of the public can be quite different from that of domain experts. Understanding how this can happen is another interesting topic that can be further investigated.

6 Conclusion

In this paper, we proposed a tool for visualizing heterogeneous semantic networks obtained from multiple data sources. The modeling of the metadata for the entities and relationships contained in semantic networks and their mappings are described. For easy analysis of the integrated network and compare their differences, we have provided a visualization interface using TouchGraph API. A case study on two semantic networks obtained from TKB and Wikipedia is reported to illustrate the differences in the understanding of terrorism related information from the public and the expert domain.

The future work for this visualization tool is to embed the system with functionality to query the graph using faceted search technique [15]. Faceted search is basically a method for refining search results by categories. For example, given a library of terrorism from our database, faceted search will enable user to pare down the results of his search using attributes such as location of incident, date of event, terrorist's nationality and so on. Thus, this method will allow the user to browse and navigate the information that they want to search for.

As for the other future works, we will focus on minimizing the job of manual entity matching as well as further enhancing the interface with more helpful features. Zooming network with a fish-eye view for complex network, back/forward option for users to retrace their steps during browsing, load/save the current network view for selected node are some possible enhancements.

7 Acknowledgement

This work was supported by A*STAR Public Sector R&D, Singapore, Project Number 062 101 0031.

References

1. H. Alani. Tgviztab: An ontology visualization extension for protege. In *Proc. of Knowledge Capture (K-Cap'03), Workshop on Visualization Information in Knowledge Engineering*, Sanibel Island, Florida, 2003.
2. C. Chen. Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, pages 401–420, 1999.
3. P. Ciccarese, S. Mazzocchi, F. Ferrazzia, and L. Sacchia. Genius: a new tool for gene networks visualization. In *Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP) Proceedings*, pages 107–111, 2004.
4. P. W. Eklund, N. Roberts, and S. Green. Ontorama: Browsing an rdf ontology using a hyperbolic-like browser. In *Proc. of the First International Symposium on CyberWorlds (CW2002) Theory and Practices*, pages 405–411, Seattle, Washington, 2002.
5. J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *Proc. IEEE Symposium on Information Visualization*, Minneapolis, MN, USA, 2005.
6. I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.
7. K. T. J. Jankun and L. M. Kwan. Moiregraphs: Radial focus plus context visualization and interaction for graphs with visual nodes. In *Proc. of IEEE Symposium on Information Visualization*, pages 20–21, Seattle, Washington, 2003.
8. A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou. Ontology visualization methods—a survey. *ACM Comput. Surv.*, 39(4):10, 2007.
9. A. C. Lear. Xml seen as integral to application integration. *IT Professional*, 1(5):12–16, 1999.
10. Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169, 2004.
11. OntoViz. http://protegewiki.stanford.edu/index.php/OntoViz.
12. M. Storey, R. Lintern, and N. Ernst. Visualization and protege. *7th International Protege Conference*, 2004.
13. B. Suh and B. B. Bederson. Ozone: A zoomable interface for navigating ontology information. In *Proc. of Advanced Visual Interfaces*. ACM, 2002.
14. Touchgraph. http://www.touchgraph.com/.
15. K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 2003. ACM.